Biometrics 61, 982–991
December 2005
DOI: 10.1111/j.1541-0420.2005.00364.x

A Case-Cohort Design for Assessing Covariate Effects in Longitudinal Studies

Ruth M. Pfeiffer,[1,*] Louise Ryan,[2] Augusto Litonjua,[3] and David Pee[4]

[1]Biostatistics Branch, National Cancer Institute, DCEG, EPS/8030, Bethesda, Maryland 20892-7244, U.S.A.
[2]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A.
[3]Channing Laboratory, Harvard Medical School, Boston, Massachusetts 02115, U.S.A.
[4]Information Management Services, Inc., Rockville, Maryland 20852, U.S.A.
[*]*email:* pfeiffer@mail.nih.gov

Summary. The case-cohort design for longitudinal data consists of a subcohort sampled at the beginning of the study that is followed repeatedly over time, and a case sample that is ascertained through the course of the study. Although some members in the subcohort may experience events over the study period, we refer to it as the "control-cohort." The case sample is a random sample of subjects not in the control-cohort, who have experienced at least one event during the study period. Different correlations among repeated observations on the same individual are accommodated by a two-level random-effects model. This design allows consistent estimation of all parameters estimable in a cohort design and is a cost-effective way to study the effects of covariates on repeated observations of relatively rare binary outcomes when exposure assessment is expensive. It is an extension of the case-cohort design (Prentice, 1986, *Biometrika* **73,** 1–11) and the bidirectional case-crossover design (Navidi, 1998, *Biometrics* **54,** 596–605). A simulation study compares the efficiency of the longitudinal case-cohort design to a full cohort analysis, and we find that in certain situations up to 90% efficiency can be obtained with half the sample size required for a full cohort analysis. A bootstrap method is presented that permits testing for intra-subject homogeneity in the presence of unidentifiable nuisance parameters in the two-level random-effects model. As an illustration we apply the design to data from an ongoing study of childhood asthma.

Key words: Biased sampling; Cohort study; Correlated binary data; Nested random-effects model.

1. Introduction

In some epidemiologic investigations, the most expensive part of the study is not in ascertaining subjects, but in measuring their exposures and predictors of interest. This is the case in the example that motivated this article, the Home Allergen Study of childhood asthma, consisting of approximately 500 children, who were recruited as newborn infants, and are followed prospectively (for details see Gold et al., 1999). The main objective of the study is to assess the role of immune function in impacting asthma risk directly, and in modifying the risk associated with various environmental exposures. Immune function is generally assessed through biomarkers, such as cytokine proliferation levels, measured in cord blood samples that can be archived for later assaying. Because the assessment of these biomarkers is very expensive and labor intensive, the development of cost-effective subsampling strategies is highly desirable.

We propose a case-cohort design for longitudinal data as a cost-effective way to study the effects of covariates on repeated observations of relatively rare binary outcomes. Our design entails choosing a random subcohort at the beginning of the study and following the subjects in this subcohort repeatedly over time. Although some members in this random subcohort may in fact experience the events of interest over the course of the study, we refer to them as the "control-cohort" for simplicity. In addition to the control-cohort we obtain a "case sample," a random sample of subjects not in the control-cohort, who have already experienced at least one event during the study period.

The case-cohort design introduced by Prentice (1986), closely related to designs proposed earlier by Kupper, McMichael, and Spirtas (1975) and Miettinen (1982), also involves collecting covariate data only for cases experiencing the event of interest in a cohort and for members of a randomly selected subcohort. This design is based on the observation that for rare outcomes the efficiency of relative risk estimation for the Cox proportional hazards model is largely constrained by the total number of cases. Aside from our extension to longitudinal data, we limit our observations to a random subset of the cases, whereas the standard case-cohort design observes all cases occurring in the cohort. As the margins, that is, the number of cases and controls that is being analyzed, are not fixed, and some of the subjects who were previously ascertained as controls can actually experience events later on in the study, we can estimate the absolute risk of the event of interest in the population at large, similar to the nested case-control approach (Langholz and Borgan, 1997).

The longitudinal case-cohort design also extends the bidirectional case-crossover design introduced by Navidi (1998). This design generalizes the standard case-crossover design (Maclure, 1991), where only cases with a single failure time are ascertained from a cohort. In the original case-crossover design each subject is considered to be a stratum in a case-control study, the failure times represent the cases, and the other times are the controls. Inference proceeds by conditional logistic regression with exposures at the case times compared with exposures at the control times. Control time can only be time that precedes the event, which can lead to confounding by time trends in the exposure. The bidirectional case-crossover design circumvents this problem by comparing exposures at failure with exposures before and after failure. In addition, multiple failure times can be dealt with by conditioning on the exact number of events a person experiences in the study period, which limits one to estimating only time-varying covariates. By introducing a control-cohort and allowing for different correlations between the observations taken on the same individual, we can estimate the effects of exposures that are constant for a subject as well as the effects of time-varying covariates. In fact, our design allows for estimation of all parameters that would be estimable based on a longitudinal cohort design.

In Section 2, we derive the likelihood for the proposed longitudinal case-cohort design. In Section 3, we propose a score test to test for the necessity to model intra-individual correlations. In Section 4, we assess the performance of the estimates of covariate effects in a simulation study and compare their efficiency to those of estimates based on a full cohort analysis. Data from the ongoing Home Allergen Study are used to illustrate our proposed design in Section 5. We conclude with a discussion of our results in Section 6.

2. Data and Model

Assume that individuals are ascertained during a fixed time period from a cohort at risk, and that events can occur at any of the fixed T time points $t_1, \ldots, t_T$. Let Y_{it}, $i = 1, \ldots, n$, $t = 1, \ldots, T$, represent the binary outcome of individual i at time t, with $Y_{it} = 1$ if the ith individual experienced an event at time t and $Y_{it} = 0$ otherwise. Covariates are denoted by X_{it} and may include components that are either time varying or independent of time.

The data consist of two parts: a control-cohort, that is a random sample of size n_0 of the population chosen at $t = 0$ and followed forward in time, and a so-called "case sample." The latter is a random sample of individuals who were not selected as part of the control-cohort, and who experienced at least one event during the course of the study. Once a case is ascertained into the study, the exposure history up to the time of the event is reconstructed, and from the event time on, the case is followed forward in time. This is a reasonable sampling scheme if specimens are stored at baseline and are available for later evaluation.

2.1 *The Random-Effects Model*

The probability p_{it} that individual i in the cohort has an event at time t follows the logistic model

$$\begin{aligned}\operatorname{logit}(p_{it}) &= \operatorname{logit} P(Y_{it} = 1 \mid a_i, g_{it}, X_{it}) \\ &= \mu + \sigma_a a_i + \sigma_g g_{it} + \beta X_{it}. \end{aligned} \tag{1}$$

The random effect a_i models individual-specific effects that are common to all time points while the random effects g_{it}'s allow the observations for each individual to be differently correlated. The a_i are assumed to be independent and identically distributed (i.i.d.) with $E(a_i) = 0$ and $\operatorname{var}(a_i) = 1$. The a_i are also assumed to be independent of the g_{it}'s, which have mean 0, variance 1, and are serially correlated for the ith individual. For many longitudinal settings, it might be desirable to assume that the g's arise from a stationary AR(1) process (see, for example, Brockwell and Davies, 1991, p. 79). This means that $g_{it} = \gamma g_{i(t-1)} + \epsilon_{it}$, where the ϵ_{it}'s are i.i.d. and have an $N(0, \sigma_\epsilon^2)$ distribution, $|\gamma| \le 1$, and $g_{i0} \sim N(0, \sigma_\epsilon^2/(1-\gamma^2))$. Under this model, and for equally spaced time points, the g's have a normal distribution with mean 0 and, using $\sigma_g^2 = \sigma_\epsilon^2/(1-\gamma^2)$ in (1), correlation matrix

$$\Sigma_g = \begin{pmatrix} 1.0 & \gamma & \gamma^2 & \gamma^3 & \cdots & \gamma^T \\ \gamma & 1.0 & \gamma & \gamma^2 & \cdots & \gamma^{T-1} \\ \gamma^2 & \gamma & 1.0 & \gamma & \cdots & \gamma^{T-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma^T & \gamma^{T-1} & \gamma^{T-2} & \gamma^{T-3} & \cdots & 1.0 \end{pmatrix}. \tag{2}$$

This parameterization also allows for different numbers of observation for different individuals without affecting the dimensionality of the parameter space. For irregularly spaced time points the correlation matrix (2) can be adjusted accordingly. The autoregressive model allows for a rich dependence structure: large positive values of γ induce gradual changes in the probability of an event over time, while for γ close to 0 the observations of a subject are nearly independent. Stationarity is a desirable property, as it guarantees that all the g_{it}'s have the same variance.

Model (1) has a two-level structure and was used by Pfeiffer, Gail, and Pee (2001) to model family data, with a covariance structure determined from the relationships of the family members. It is an extension of the widely used random-effects model that allows for a cluster-specific intercept a_i, but assumes that the Y_{it}'s are conditionally independent, given a_i and the measured X_{it} (see, for example, Neuhaus, Kalbfleisch, and Hauck, 1991). Diggle, Liang, and Zeger (1994, Chapter 9) presented similar generalized linear mixed models for exponential families with canonical links. Albert et al. (2002) used a Gaussian process with mean 0 and an exponential covariance structure, $\operatorname{Cov}(g_{it}, g_{it'}) = \sigma_g^2 \exp(-\theta\,|t - t'|)$, in an informative missing data model to allow for different correlations among observations on the same individual.

The marginal probability of the response for the ith individual under the logistic model (1) is

$$P(Y_{i1}, Y_{i2}, \ldots, Y_{iT} \mid X_{i1}, \ldots, X_{iT}) = \int \prod_t p_{it}^{y_{it}} q_{it}^{1-y_{it}}\, dF(a, g), \tag{3}$$

where $q_{it} = 1 - p_{it}$.

3. Estimation and Inference

3.1 *Derivation of the Scores*

Each subject in the control-cohort is sampled randomly from the cohort and contributes the probability (3) to the likelihood. The case sample on the other hand is comprised of

subjects who experienced at least one event during the course of the study. Inference must appropriately account for this sampling mechanism. As discussed by Zhao and Lipsitz (1992) in the context of two-stage designs, there are several possible approaches. We extend the conditional-likelihood approach, which involves specifying the conditional distribution of the observed data, given that the subject was ascertained. Let $Y_{i\cdot} = \sum_{t=1}^{T} Y_{it}$ denote the total number of events individual i experiences during the study period. By applying Bayes' theorem, the conditional distribution for such an ascertained case i can be written as

$$\begin{aligned} &P(Y_{i1}, Y_{i2}, \ldots, Y_{iT} \,|\, X_{i1}, \ldots, X_{iT}, Y_{i\cdot} \geq 1) \\ &= \frac{P(Y_{i1}, Y_{i2}, \ldots, Y_{iT}, Y_{i\cdot} > 0 \,|\, X_{i1}, \ldots, X_{iT})}{P(Y_{i\cdot} \geq 1 \,|\, X_{i1}, \ldots, X_{iT})} \\ &= \frac{P(Y_{i1}, Y_{i2}, \ldots, Y_{iT}, Y_{i\cdot} > 0 \,|\, X_{i1}, \ldots, X_{iT})}{1 - P(Y_{i\cdot} = 0 \,|\, X_{i1}, \ldots, X_{iT})} \\ &= \frac{\int \prod_{t=1}^{T} p_{it}^{y_{it}} q_{it}^{1-y_{it}} \; dF(a,g)}{1 - \int \prod_{t=1}^{T} q_{it} \; dF(a,g)}. \end{aligned} \tag{4}$$

The case-cohort thus arises from a truncated mixed distribution. The probability that a case is sampled does not depend on the exact number of events a case experienced, but only on whether the person had at least one event.

Combining the scores for the control-cohort and for the case-cohort yields the log likelihood

$$\begin{aligned} \ln l(Y;\boldsymbol{\theta}) = &\sum_{i=1}^{n_0} \ln \left[\int \prod_{t=1}^{T} p_{it}^{y_{it}} q_{it}^{1-y_{it}} \; dF(a,g) \right] \\ &+ \sum_{i=1}^{n_1} \ln \left[\frac{\int \prod_{t=1}^{T} p_{it}^{y_{it}} q_{it}^{1-y_{it}} \, dF(a,g)}{1 - \int \prod_{t=1}^{T} q_{it} \, dF(a,g)} \right], \end{aligned} \tag{5}$$

where $\boldsymbol{\theta} = (\beta,\ \sigma_a^2,\ \gamma,\ \sigma_g^2)$, and $p_{it} = p_{it}(\boldsymbol{\theta},\ X_{it})$. Conceptually, this likelihood can be viewed as the combined likelihood from two separate studies. Suppose that the original cohort is split randomly into two parts. The first part gives rise to the control-cohort, and the case-cohort is a random sample of the cases of the second part. Combining both likelihoods yields (5).

To derive the asymptotic distribution of the estimates, denote the scores for the control-cohort by

$$U_i(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}'} \ln \left[\int \prod_{t=1}^{T} p_{it}^{y_{it}} q_{it}^{1-y_{it}} \; dF(a,g) \right]$$

and the scores for the case-cohort by

$$U_i^{+}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}'} \ln \left[\int \prod_{t=1}^{T} p_{it}^{y_{it}} q_{it}^{1-y_{it}} dF(a,g) \Big/ \left(1 - \int \prod_{t=1}^{T} q_{it} \; dF(a,g) \right) \right].$$

The estimator $\hat{\boldsymbol{\theta}}$ then solves

$$U(\boldsymbol{\theta}) = \sum_{i=1}^{n_0} U_i(\boldsymbol{\theta}) + \sum_{i=1}^{n_1} U_i^{+}(\boldsymbol{\theta}) = 0.$$

Consistency of the estimates follows as $E_Y U_i(\boldsymbol{\theta}) = 0$ and $E_Y U_i^{+}(\boldsymbol{\theta}) = E_{Y\cdot} E_{Y\,|\,Y\cdot} U_i(\boldsymbol{\theta}) = 0$ and thus the expected value of the combined scores is 0. The terms U_i and U_i^{+} are independent but not identically distributed, thus we use the Lindeberg–Feller theorem to derive asymptotic normality of the estimates. As $n_0, n_1 \to \infty$ such $n_1/n_0 \to \phi$, using standard Taylor expansion,

$$\begin{aligned} n_0^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = &-n_0 \left[\sum_{i=1}^{n_0} \frac{\partial}{\partial \boldsymbol{\theta}'} U_i(\boldsymbol{\theta}) + \sum_{j=1}^{n_1} \frac{\partial}{\partial \boldsymbol{\theta}'} U_j^{+}(\boldsymbol{\theta}) \right]^{-1} \\ &\times n_0^{-1/2} \left[\sum_{i=1}^{n_0} U_i(\boldsymbol{\theta}) + \sum_{j=1}^{n_1} U_j^{+}(\boldsymbol{\theta}) \right] \\ &+ o_p(1) \to N(0, A^{-1} B (A^{-1})^{\mathrm{T}}). \end{aligned}$$

The pieces of the asymptotic variance are $A = E[\frac{\partial}{\partial \boldsymbol{\theta}'} U(\theta)] + \phi E[\frac{\partial}{\partial \boldsymbol{\theta}'} U^{+}(\theta)]$, estimated by

$$\hat{A} = \frac{1}{n_0} \left[\sum_{i=1}^{n_0} \frac{\partial}{\partial \boldsymbol{\theta}'} U_i(\boldsymbol{\theta}) + \sum_{i=1}^{n_1} \frac{\partial}{\partial \boldsymbol{\theta}'} U_i^{+}(\boldsymbol{\theta}) \right]$$

and $B = E[U(\boldsymbol{\theta}) U'(\boldsymbol{\theta})] + \phi E[U^{+}(\theta) U^{+'}(\boldsymbol{\theta})]$, estimated by

$$\hat{B} = \frac{1}{n_0} \left[\sum_{i=1}^{n_0} U_i(\boldsymbol{\theta}) U_i'(\boldsymbol{\theta}) + \sum_{j=1}^{n_1} U_j^{+}(\boldsymbol{\theta}) U_j^{+'}(\boldsymbol{\theta}) \right]$$

evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

We will discuss computational issues relating to our model after making some general comments. To derive the probability distribution for the cases in the sample, we conditioned on the exact ascertainment, namely that a case in the case-cohort experienced at least one event. Alternatively, one could condition on a slightly stronger condition, the exact number of events, and find $P(Y_{i1}, Y_{i2}, \ldots, Y_{iT} \mid X_{i1}, \ldots, X_{iT}, Y_{i\cdot})$ instead of $P(Y_{i1}, Y_{i2}, \ldots, Y_{iT} \mid X_{i1}, \ldots, X_{iT}, Y_{i\cdot} \geq 1)$. In the absence of the random effects g_{it} (i.e., $g_{it} = 0$ for all i, t), this would lead to standard conditional logistic regression, as $Y_{i\cdot}$ is the sufficient statistic for the individual-specific intercept $\mu + a_i$,

$$\begin{aligned} &P(Y_{i1}, Y_{i2}, \ldots, Y_{iT} \,|\, X_{i1}, \ldots, X_{iT}, Y_{i\cdot}) \\ &= \frac{P(Y_{i1}, Y_{i2}, \ldots, Y_{iT}, Y_{i\cdot} \,|\, X_{i1}, \ldots, X_{iT})}{P(Y_{i\cdot} \,|\, X_{i1}, \ldots, X_{iT})} \\ &= \frac{\exp\left(\beta \sum_{t_j \in A_i} X_{it_j} \right)}{\sum_{S \in D_{Y_{i\cdot}}} \exp\left(\beta \sum_{t_j \in A_i} X_{it_j} \right)}, \end{aligned} \tag{6}$$

where A_i is the set of times at which failures occur. In the presence of the g_{it}'s no such simplification occurs, as can be seen from (4). In addition, evaluation of (6) will be computationally more complicated than expression (4), as there are $T!/[Y_{i.}!\,(T - Y_{i.})!]$ summands in the denominator.

In the bidirectional case-crossover design (Navidi, 1998), the likelihood contribution of an individual corresponds to (6), and is based on the assumption that the individual failure times are independent, given the covariates. Note also that based on the conditional probability (6), only individuals who had events at some but not all time points contribute to the likelihood, while everybody contributes based on the truncated version (4). O'Neill and Barry (1995) compare truncated logistic regression with conditional logistic regression and favor the former because it yields more efficient estimates and allows for estimation of cluster-level effects. However, they use the rather strong assumption that the logistic regression intercept is the same from cluster to cluster. This assumption is relaxed in our model by introducing the cluster-level random effect a. By dealing with truncated logistic regression instead of conditional logistic regression, it is in principle possible to estimate μ and σ_a from the case-cohort alone as these parameters do not cancel out of the likelihood (4) even when all the g's are 0. One would expect a case-only sample to contain very little information on these parameters however, which makes their estimation unstable. By adding the control-cohort we stabilize estimates for μ and σ_a.

3.2 *Numerical Methods for the Random-Effects Model*

Estimation of the parameters in our model shares many of the computational difficulties associated with generalized linear mixed models (see, for example, Breslow and Clayton, 1993), as evaluating high-dimensional integrals is required.

We estimate the parameters of the likelihood by direct maximization, and evaluate the integrals in (5) by Monte Carlo integration (see Tanner, 1993, p. 30). For the ith individual in the data set, we draw N independent, identically distributed samples a_{ik}, $k = 1, \ldots, N$, from a standard normal distribution, and independent, identically distributed $\mathbf{g}_k^* = (g_{ik1}^*, \ldots, g_{ikT}^*)$, $k = 1, \ldots, N$, from a multivariate normal distribution with mean vector 0 and identity correlation matrix. For each iteration in the maximization, we then compute $\mathbf{g}_k = \Sigma_g^{1/2}\mathbf{g}_k^*$ to obtain random effects with correlation matrix (2) and use the approximation

$$\begin{aligned}&\int \prod_{t=1}^{T} p_{it}(a_i, g_{i1}, \ldots, g_{iT})^{y_{it}} q_{it}(a_i, g_{i1}, \ldots, g_{iT})^{1-y_{it}}\; dF(a, g)\\ &\quad\approx \frac{1}{N}\sum_{k}^{N}\prod_{t=1}^{T} p_{it}(a_{ik}, g_{ik1}, \ldots, g_{ikT})^{y_{it}}\\ &\qquad\times q_{it}(a_{ik}, g_{ik1}, \ldots, g_{ikT})^{1-y_{it}}\end{aligned}$$

for individuals sampled into the control-cohort. Integrals in the conditional likelihood (4) for individuals sampled into the case-cohort are evaluated the same way, with the same Monte Carlo sample used for the numerator and denominator to ensure that the conditional likelihood is smooth in β. Different individuals are evaluated using independent Monte Carlo samples. The advantage of Monte Carlo integration over Gaussian quadrature is that required computations increase only linearly with the dimension of the integral, while the numerical effort for Gaussian quadrature increases exponentially with the dimension of the integral. For the simulation study we chose $N = 500$ and for the data example $N = 1000$.

Another strategy to evaluate the integrals that was discussed by Raudenbush, Yang, and Yosef (2000) would be the use of higher-order Laplace approximations. Raudenbush et al. showed through simulations that approximating the log likelihood by a sixth-order approximation was as accurate as quadrature but considerably faster. Even though this approach seems very promising, it is rather complicated in our situation, as the likelihood is composed of two different pieces. For our purposes Monte Carlo integration is adequately accurate, has acceptable computational speed, and is very easy to implement.

3.3 *A Score Test for Residual Correlation*

We derive a score test to formally test $\mathrm{H}_0 : \sigma_g = 0$, that is, to test whether the observations of an individual are independent, given the covariates and the individual-specific intercept a_i. Tests of $\sigma_g = 0$ for this model can also be used as partial goodness-of-fit tests.

The derivations follow a strategy developed by Liang (1987) and extended by Commenges and JacqminGadda (1997). First, we reparameterize p_{ij} using $\nu = \sigma_g^2$ as

$$\begin{aligned}\operatorname{logit}(p_{ij}) &= \operatorname{logit} P(Y_{ij} = 1 \,|\, a_i, g_{ij}, X_{ij})\\ &= \mu + \sigma_a a_i + \sqrt{\nu} g_{ij} + \beta X_{ij}.\end{aligned}$$

The score statistic for ν for the ith individual in the control-cohort evaluated at the null hypothesis is given by

$$\begin{aligned}&S_\nu(Y_j, \theta, X_i)\\ &\quad= (\partial/\partial\nu)\ln\{P(Y_{j1}, Y_{j2}, \ldots, Y_{jT}\,|X_{j1}, \ldots, X_{jT})\}|_{\,\nu=0},\end{aligned}$$

and the score for $\boldsymbol{\theta} = (\mu, \sigma_a^2, \beta, \gamma)$ at $\nu = 0$ is

$$\begin{aligned}&S_\theta(Y_j, \theta, X_j)\\ &\quad= (\partial/\partial\boldsymbol{\theta})\ln\{P(Y_{j1}, Y_{j2}, \ldots, Y_{jT}\,|X_{j1}, \ldots, X_{jT})\}|_{\,\nu=0},\end{aligned}$$

while for an individual j in the case-cohort we have

$$\begin{aligned}&S_\nu^+(Y_i, \theta, X_i)\\ &\quad= (\partial/\partial\nu)\ln\{P(Y_{i1}, Y_{i2}, \ldots, Y_{iT}\,|X_{i1}, \ldots, X_{iT}, Y_{i.} \geq 1)\}|_{\,\nu=0},\end{aligned}$$

and

$$\begin{aligned}&S_\theta^+(Y_i, \theta, X_i)\\ &\quad= (\partial/\partial\boldsymbol{\theta})\ln\{P(Y_{i1}, Y_{i2}, \ldots, Y_{iT}\,|X_{i1}, \ldots, X_{iT}, Y_{i.} \geq 1)\}|_{\,\nu=0}.\end{aligned}$$

Omitting the subject index for simplicity, S_ν^+ and S_ν are found by applying l'Hospital's rule to be

$$S_\nu^+(Y,\theta,X)|_{\nu=0} = \frac{\int \prod p_k^{y_k} q_k^{1-y_k} \left\{ \sum_j \left[(y_j-p_j)^2 - p_j q_j\right] + \sum_j \sum_{k\neq j} E g_k g_j (y_j-p_j)(y_k-p_k) \right\} dF_a}{2\int \prod p_k^{y_k} q_k^{1-y_k} dF_a} - \frac{\int \prod_l q_l \left\{ \sum_i p_i q_i - \sum_k \sum_{j\neq k} p_k p_j E g_k g_j \right\}}{2\left(1 - \int \prod q_i \, dF_a\right)}$$

and

$$S_\nu(Y,\theta,X)|_{\nu=0} = \frac{\int \prod p_k^{y_k} q_k^{1-y_k} \left\{ \sum_j \left[(y_j-p_j)^2 - p_j q_j\right] + \sum_j \sum_{k\neq j} E g_k g_j (y_j-p_j)(y_k-p_k) \right\} dF_a}{2\int \prod p_k^{y_k} q_k^{1-y_k} \, dF_a}.$$

The maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ under the hypothesis that $\nu = 0$ are computed numerically, as well as the derivatives of S_θ and S_θ^+. To fit the model with only the random effect a_i is considerably simpler, as in this case the likelihood (5) involves only one-dimensional integrals, that can be evaluated using Gaussian quadrature, for example.

Using standard likelihood theory, the combined score test statistic for the control-cohort and case sample is

$$T_{\hat\theta} = \left(\frac{n_0+n_1}{n_0+n_1-d}\right)^{-1/2} \times \left\{ \sum_{i=1}^{n_0} S_\nu(Y_i,\hat{\boldsymbol{\theta}}) + \sum_{j=1}^{n_1} S_\nu^+(Y_j,\hat{\boldsymbol{\theta}}) \right\} \Big/ \sqrt{I}, \quad (7)$$

where d denotes the number of components of $\boldsymbol{\theta}$ and the leading term is a correction factor for the degrees of freedom. The denominator of the score test is given by

$$\begin{aligned} I = {} & \sum_{i=1}^{n_0} I_{\nu\nu i}(\hat{\boldsymbol{\theta}}) + \sum_{j=1}^{n_1} I^+_{\nu\nu j}(\hat{\boldsymbol{\theta}}) \\ & - \left\{ \sum_{i=1}^{n_0} I_{\nu\theta i}(\hat{\boldsymbol{\theta}}) + \sum_{j=1}^{n_1} I^+_{\nu\theta j}(\hat{\boldsymbol{\theta}}) \right\} \\ & \times \left\{ \sum_{i=1}^{n_0} I_{\theta\theta i}(\hat{\boldsymbol{\theta}}) + \sum_{j=1}^{n_1} I^+_{\theta\theta j}(\hat{\boldsymbol{\theta}}) \right\}^{-1} \\ & \times \left\{ \sum_{i=1}^{n_0} I'_{\nu\theta i}(\hat{\boldsymbol{\theta}}) + \sum_{j=1}^{n_1} I^{+'}_{\nu\theta i}(\hat{\boldsymbol{\theta}}) \right\} \end{aligned}$$

with

$$\begin{aligned} I_{\nu\nu i} &= E\left[S_\nu^2(Y_i,\boldsymbol{\theta}) \,|\, \nu = 0\right], \\ I_{\nu\theta i} &= E\left[S_\nu(Y_i,\boldsymbol{\theta}) S'_\theta(Y_i,\boldsymbol{\theta}) \,|\, \nu = 0\right], \\ I_{\theta\theta i} &= E\left[S_\theta(Y_i,\boldsymbol{\theta}) S'_\theta(Y_i,\boldsymbol{\theta}) \,|\, \nu = 0\right], \end{aligned}$$

for the control-cohort, and $I^+_{\nu\nu i}$, $I^+_{\nu\theta i}$, and $I^+_{\theta\theta i}$ defined similarly for the case-cohort. In our problem the expectations are too difficult to compute, and the expected Fisher information is replaced by the observed Fisher information, for example, $I_{\nu\nu i} = S_\nu^2(Y_i,\, \theta)|_{\nu=0}$.

A difficulty is that under H_0 the parameters in the distribution of the g'_{it}'s are not identifiable, yet the test statistic depends on them through Eg_ig_k. In our example, Eg_ig_k depends on the parameter γ in the autoregressive model. While Commenges and JacqminGadda (1997) circumvent this problem by assuming the covariance matrix of the random effects is known, we modify an approach suggested by Davies (1977, 1987) for this situation. We now let $\boldsymbol{\theta} = (\mu,\, \sigma_a^2,\, \beta)$ denote the parameters that are identifiable under H_0. Following Davies (1977), we rewrite the test statistic in (7) as $T_\theta(\gamma)$, and when γ is unknown, replace it with

$$T_\theta^* = \sup\{T_\theta(\gamma) : l \le \gamma \le u\}$$

where $[l,\, u]$ is the range of possible values of γ, in our setting $[l,\, u] = [-1,\, 1]$. Under the assumption that $T_\theta(\gamma)$ is continuous on $[l,\, u]$ with a continuous derivative except possibly for a finite number of jumps, and $T_\theta(\gamma)$ has a normal distribution for every value of γ, that is, $T_\theta(\gamma)$ is a Gaussian process in γ, Davies derived the following upper bound as an approximation of the p-value of the test statistic:

$$\begin{aligned} & P(\sup T_\theta(\gamma) > c : l \le \gamma \le u) \\ & \quad \le \Phi(-c) + \exp(-1/2c^2) \int_l^u \{-\rho_{11}(\nu)\}^{1/2} \, d\nu/2\pi, \end{aligned}$$

where $\rho_{11}(\gamma) = [\partial^2\rho(\phi,\, \gamma)/\partial\phi^2]_{\phi=\gamma}$ and $\rho(\phi,\, \gamma) = \mathrm{corr}\{T_\theta(\gamma),\, T_\theta(\phi)\}$ denote the autocorrelation function of $T_\theta(\gamma)$.

Instead of the above upper bound, we propose a parametric bootstrap to obtain an approximate p-value of the test statistic. The bootstrap for the longitudinal case-cohort design consists of the following steps:

1. Generate a bootstrap sample for the control-cohort $Y_1^b, \ldots, Y_{n_0}^b$ and a bootstrap sample for the case sample $Y_1^b, \ldots, Y_{n_1}^b$ conditional on the covariates under H_0 with $\boldsymbol{\theta} = (\mu,\, \sigma_a^2,\, \beta)$ replaced by $\hat{\boldsymbol{\theta}}$ estimated from the original data under H_0. That is, given the covariates X_{it} generate outcomes from $\mathrm{logit}\, P(\, Y_{it} = 1 \mid a_i,\, g_{it},\, X_{it}) = \mu + \sigma_a a_i + \beta X_{\,it}$. Note that each bootstrap sample for the case sample needs to satisfy $Y_{\cdot}^b \ge 1$, thus several draws for a given covariate vector may be necessary to ensure that the sample comes from the conditional distribution.
2. Estimate $\tilde{\boldsymbol{\theta}}_b$ for the bootstrap sample under H_0.
3. Compute $T^*_{\tilde\theta} = \sup T_{\tilde\theta}(\gamma)$ for $\gamma \in [\, -1,\, -0.95,\, ,\, -0.9,\, -0.85, \ldots,\, 0.9,\, 0.95,\, 1.0]$.
4. Repeat steps 1–3 B times to obtain the bootstrap distribution function of T_θ^* under H_0.

We then use the jth-order statistic of the B bootstrap replications as an estimate of the $j/(B+1)$st quantile. The test is necessarily a one-sided test since the parameter value specified by the null hypothesis is on the boundary of the parameter space.

For comparison, the test statistic to test for residual correlation using the whole cohort of size n is $T_{\hat{\theta}} = \sum_{i=1}^{n} S_\nu(Y_i, \hat{\theta})/(I)^{1/2}$, with $I = \sum_{i=1}^{n} I_{\nu\nu i}(\hat{\theta}) - \{\sum_{i=1}^{n} I_{\nu\theta i}(\hat{\theta})\}\{\sum_{i=1}^{n} I_{\theta\theta i}(\hat{\theta})\}^{-1}\{\sum_{i=1}^{n} I'_{\nu\theta i}(\hat{\theta})\}$. To find the distribution of $T_\theta(\gamma)$ when γ is unknown, step one of the bootstrap simplifies to resampling from the whole cohort for computing $\boldsymbol{\theta}_b$.

4. Simulation Study

4.1 *Estimation Results for the Longitudinal Case-Cohort Design*

We used simulated data to assess behavior and efficiency of the estimates for the longitudinal case-cohort design. In the simulated data sets we assumed equally spaced observation times $t_1, \ldots, t_6$ for all individuals. In model (1) we let $\beta = 1$, with a time-varying covariate X_t from a Bernoulli distribution, that is, $X_t \in \{0, 1\}$ with $p = 0.5$, and various choices for values of μ, σ_a, γ, and σ_g. The random effects a_i were normally distributed with mean 0 and variance 1, and the random effects g were multivariate normal with mean 0 and correlation matrix (2). The intercept parameters were $\mu = -2, -3, -4$, and -5. To put these values into perspective, note that among unexposed individuals, an intercept of $\mu = -2$ corresponds to a risk of disease of 119 per 1000, when there are no random effects in the model, and $\mu = -3$ to 47 per 1000. For $\sigma_g = 1$, $\sigma_a = 1$, $\beta = 0$, and $\mu = -2$ the disease prevalence is 185 per 1000, and for $\mu = -3$ it is 92 per 1000.

We assessed the efficiency of the estimates, defined as the ratio of the mean empirical variance estimates, of the case-cohort design compared to a full cohort analysis (Table 1). For each choice of parameters, we fit model (3) to all subjects in the simulated cohort and then sampled a control-cohort and a case sample and computed estimates based on the log likelihood (5).

The longitudinal case-cohort design yielded nearly unbiased estimates for $\beta = 1$ and near nominal 95% coverage of β for likelihood-ratio-based confidence intervals (CIs) (data not shown) for each of the parameter combinations studied in Table 1. While μ was estimated without much bias and reasonable precision as well, the estimates of σ_g and σ_a were associated with large standard errors, and estimates of σ_g had much larger coefficients of variation than estimates of β. Likelihood-ratio-based confidence intervals for σ_g^2 and μ had near nominal 95% coverage, but coverage was subnominal for σ_a^2 and γ. Unreported confidence intervals based on the Wald statistic for β and σ_g had subnominal coverage. In many applications however, the main interest lies in estimation of β, and the random-effects parameters and μ will be of lesser concern.

Table 1
Comparison of efficiency of a full cohort analysis and the longitudinal case-cohort design for estimation of $\boldsymbol{\theta} = (\mu, \sigma_a^2, \beta, \sigma_g^2, \gamma)$ *based on* 100 *simulations* (*Monte Carlo sample size* 500)

μ, σ_a^2, β, σ_g^2, γ	Case-cohort analysis mean $(\hat{\mu}, \hat{\sigma}_a, \hat{\beta}, \hat{\sigma}_\epsilon^2, \gamma)$	Full cohort analysis mean $(\hat{\mu}, \hat{\sigma}_a, \hat{\beta}, \hat{\sigma}_\epsilon^2, \gamma)$
	$n_0 = 200$, $n_1 = 200$	$n = 800$
-2, 1.0, 1.0, 1.0, 0.5	-2.17, 0.94, 1.09, 2.23, 0.56	-2.07, 0.80, 1.04, 1.71, 0.59
empirical standard errors	0.49, 0.98, 0.25, 2.73, 0.26	0.35, 0.74, 0.20, 1.56, 0.26
Efficiency		**0.51, 0.57, 0.64, 0.33, 1.00**
-3, 1.0, 1.0, 1.0, 0.5	-3.22, 0.95, 1.08, 1.93, 0.57	-3.10, 0.80, 1.04, 1.58, 0.62
empirical standard errors	0.65, 0.90, 0.22, 2.42, 0.28	0.46, 0.67, 0.17, 1.33, 0.26
Efficiency		**0.50, 0.55, 0.60, 0.30, 0.86**
-2, 1.0, 1.0, 2.0, 0.5	-2.08, 0.92, 1.05, 2.62, 0.57	-1.98, 0.71, 0.99, 2.23, 0.64
empirical standard errors	0.41, 1.01, 0.24, 2.10, 0.21	0.36, 0.91, 0.21, 1.94, 0.19
Efficiency		**0.77, 0.81, 0.78, 0.85, 0.82**
-3, 1.0, 1.0, 10.0, 0.9	-3.04, 2.12, 0.99, 9.70, 0.88	-3.03, 2.11, 0.99, 9.40, 0.87
empirical standard errors	0.39, 2.57, 0.18, 3.14, 0.08	0.27, 2.50, 0.15, 2.65, 0.06
Efficiency		**0.49, 0.95, 0.71, 0.71, 0.56**
	$n_0 = 300$, $n_1 = 300$	$n = 1200$
-3, 1.0, 1.0, 5.0, 0.9	-3.03, 1.57, 1.01, 4.74, 0.87	-3.01, 1.47, 1.01, 4.63, 0.87
empirical standard errors	0.26, 1.59, 0.14, 1.49, 0.09	0.16, 1.45, 0.11, 1.39, 0.08
Efficiency		**0.38, 0.83, 0.62, 0.87, 0.79**
	$n_0 = 400$, $n_1 = 400$	$n = 1600$
-4, 1.0, 1.0, 8.0, 0.9	-3.98, 1.79, 0.98, 7.14, 0.88	-3.99, 1.72, 0.98, 7.21, 0.88
empirical standard errors	0.33, 1.97, 0.11, 1.87, 0.07	0.21, 1.92, 0.10, 1.83, 0.07
Efficiency		**0.41, 0.95, 0.83, 0.96, 1.00**
	$n_0 = 200$, $n_1 = 200$	$n = 4000$
-5, 1.0, 1.0, 1.0, 0.5	-5.20, 0.60, 1.04, 1.89, 0.61	-4.92, 0.50, 0.98, 1.34, 0.67
empirical standard errors	0.66, 0.65, 0.19, 1.53, 0.26	0.26, 0.48, 0.10, 0.67, 0.28
Efficiency		**0.16, 0.55, 0.28, 0.19, 1.16**

Table 2
Actual rejections of the score test for $H_0 : \sigma_g^2 = 0$ based on the longitudinal case-cohort design for $\alpha = 0.05$ with 100 bootstrap replications

n_0, n_1	μ, σ_a^2, β, σ_g^2, γ	Rejection/total runs
200, 200	−5, 1.0, 1.0, 0.0, 0.0	0.03 (6/231)
200, 200	−2, 1.0, 1.0, 0.0, 0.0	0.07 (27/365)
200, 200	−3, 1.0, 1.0, 2.0, 0.5	0.22 (26/116)
200, 200	−3, 1.0, 1.0, 1.0, 0.5	0.18 (21/114)
200, 200	−2, 1.0, 1.0, 1.0, 0.5	0.10 (11/112)
200, 200	−3, 1.0, 1.0, 5.0, 0.9	0.37 (42/115)
200, 200	−3, 1.0, 1.0, 10.0, 0.9	0.93 (108/116)

For the settings in Table 1 where the total case-cohort sample size was 50% of the cohort sample size, the efficiency in estimating β based on our design was at least 60%. For $\mu = -2$, $\sigma_a = 1$, $\sigma_g^2 = 2$, and $\gamma = 0.5$ the case-cohort design was 78% efficient for the estimation of β, and for $\mu = -4$, $\sigma_a = 1$, $\sigma_g^2 = 0.8$, and $\gamma = 0.9$ the estimates of β were 83% efficient compared to a full cohort design with twice the sample size.

For the situation of a rare disease, with $\mu = -5$, $\sigma_a = 1$, $\sigma_g^2 = 1$, and $\gamma = 0.5$, and a cohort size that was 10 times larger than the case-cohort sample, the efficiency of the estimates of β was 27%, which corresponds to 2.7 times the efficiency per sample.

Table 2 assesses the performance of the score test. For $\mu = -5$, $\sigma_a = 1$, and $\beta = 2$, the estimated size at $\sigma_g = 0$ of a nominal 5% level test was found to be 3%, with 95% CI (0.005, 0.05); for $\mu = -2$ and the other parameters unchanged, the estimated size was 7% with 95% CI (0.047, 0.1). The power of the score test was 0.22 for $\mu = -3$, $\sigma_a^2 = 1$, $\gamma = 0.5$ and $\sigma_g^2 = 2$ based on a sample of $n_0 = 200$ and $n_1 = 200$. For $\mu = -3$, $\sigma_a^2 = 1$, $\gamma = 0.9$, and $\sigma_g^2 = 5$ the power was 37%. When the parameters changed to $\gamma = 0.9$ and $\sigma_g^2 = 10$, the power increased to 93%. While $\sigma_g^2 = 10$ seems large, recall that $\sigma_g^2 = \sigma_\epsilon^2/(1 - \gamma^2)$, and the above values of $\sigma_g^2 = 10$ and $\gamma = 0.9$ correspond to $\sigma_\epsilon^2 = 1.9$, a reasonable value for the variance in the AR(1) process that gives rise to the g'_{it}'s. In summary, our results indicate that the score test only detects large departures from the conditional independence assumption, that is, the assumption that the observations of an individual are independent, given the random intercept a_i.

4.2 *Robustness to Misspecification of the Individual-Level Random Effects*

Although model (1) accounts for different correlations among observations from the same individual, one needs to specify the distribution of the random effects g_{it}. Hartford and Davidian (2000) investigated violations of the assumptions of normality of the random-effects distribution in nonlinear mixed-effects models via simulations, using first-order expansions and Laplace approximations to evaluate the integrals. Due to the ascertainment correction for the case-cohort, our likelihood does not fall into any of the standard mixed effects model frameworks studied by Hartford and Davidian (2000). We thus examined the robustness of the parameter estimates in model (1) against violations of the assumptions made for the individual-level random effects in a simulation study.

Starting with $g_{i0} \sim F$, we chose $g_{it} = \gamma g_{i(t-1)} + \epsilon_{it}$ where the ϵ_{it}'s were i.i.d. and had the same distribution F in model (1). We then fit the log likelihood under the assumption of a multivariate normal distribution for the g_{it}'s with correlation structure (2). We computed means over 100 simulations with $\mu = -2$, $\beta = 1$, $\sigma_a^2 = 1$, $\gamma = 0.5$, $\sigma_g^2 = 1$, and $n_0 = n_1 = 200$, for two choices of distribution F.

First, we simulated the random effects using $F = t(k)$, a t-distribution with k degrees of freedom. For $k = 2$ degrees of freedom, the estimates for the parameters in model (1) were $\hat{\beta} = 1.01(0.29)$, $\hat{\mu} = -2.07(0.53)$, $\hat{\sigma}_a^2 = 0.85(1.31)$, $\hat{\gamma} = 0.60(0.14)$, and $\hat{\sigma}_g^2 = 4.13(7.77)$. For a t-distribution with 3 degrees of freedom, we obtained $\hat{\beta} = 1.10(0.35)$ and $\hat{\mu} = -2.23(0.63)$, $\hat{\sigma}_a^2 = 1.13(1.67)$, $\hat{\gamma} = 0.52(0.28)$, and $\hat{\sigma}_g^2 = 2.10(4.11)$. When the number of degrees of freedom was increased to 10, we observed $\hat{\beta} = 1.09(0.40)$ and $\hat{\mu} = -2.18(0.75)$, $\hat{\sigma}_a^2 = 1.11(2.03)$, $\hat{\gamma} = 0.54(0.27)$, and $\hat{\sigma}_g^2 = 2.34(5.99)$.

To study the behavior of the model when the random-effects distribution is skewed, we let F be an exponential distribution that we centered by its mean. For an exponential distribution with parameter 1, the estimates were $\hat{\beta} = 1.11(0.46)$ and $\hat{\mu} = -2.30(0.90)$, $\hat{\sigma}_a^2 = 1.24(3.62)$, $\hat{\gamma} = 0.55(0.24)$, and $\hat{\sigma}_g^2 = 3.11(11.18)$. For an exponential distribution with parameter 0.2, the estimates were $\hat{\beta} = 1.04(0.28)$ and $\hat{\mu} = -2.18(0.53)$, $\hat{\sigma}_a^2 = 0.93(1.00)$, $\hat{\gamma} = 0.57(0.25)$, and $\hat{\sigma}_g^2 = 2.20(3.25)$.

For one simulation we used g'_{it}'s from a Gaussian AR(1) process in model (1) to generate the data but fitted a logistic regression model involving only the parameters β, μ, and σ_a^2. The estimates obtained in this situation were $\hat{\beta} = 0.88(0.11)$, $\hat{\mu} = -1.70(0.12)$, and $\hat{\sigma}_a^2 = 1.01(0.19)$.

Our simulation study thus showed that ignoring the individual-level random effects completely resulted in a similar bias in the parameter estimates as when the individual-level random effects came from a heavily skewed distribution but were modeled as multivariate normal. Misspecification of the distribution of underlying random effects in our design therefore does not lead to serious bias in the parameters in most situations that are of practical relevance.

5. Example

We applied the case-cohort design to data from the Home Allergen Study. We used a sample of 398 children with complete information for the 5 years of follow-up for our example. The outcomes were the presence/absence of wheeze or asthma at prespecified observation times for each child in the study. The follow-up began when children were born, with closer spaced observation times when the children were young, for example, every 2 months during the first 2 years of life, every 6 months from age 2 to age 5 years, and more spread as the children grew older, for example, once every year from age 5 years on. All the children in our data set were born between 1994 and 1996. We used data on the children between ages 0 and 5 years, with 18 observations on each child, corresponding to a total number of 7164 observations.

There were 716 events during the study period, and the number of events per child ranged from 0 to 12. One hundred and forty-seven (36.9%) children never experienced asthma or

wheezing, 93 (23.4%) had exactly one event, and 4 children experienced 10 or more events.

We studied the effects of the presence of cats in the house and endotoxin levels measured from dust samples taken from the main living areas of the child's home on disease risk. A case-cohort design is possible in this setting since dust samples can be collected and stored for later endotoxin assay. The covariates we considered were indicator variables $X_1 = 1$ if there was a cat in the home and 0 otherwise, and two indicators for endotoxin exposure, $X_2 = 1$ for endotoxin levels in the second tertile, $X_2 = 0$ otherwise, and $X_3 = 1$ for endotoxin levels in the highest tertile, $X_3 = 0$ otherwise. The lowest tertile for endotoxin levels was the reference group. To assess whether early childhood exposures to aeroallergens may play a role in allergic sensitization and in the development and exacerbation of asthma later in life, we included interaction terms of endotoxin levels and an indicator variable that was one for children older than 12 months in the model.

We fit model (3) to the full cohort of 398 children, and then sampled 100 children into the control-cohort and 100 children who had reported wheezing at least once into the case-cohort, and computed estimates based on the case-cohort design. The g'_{it}'s were modeled using an AR(1) process with correlation structure (2), taking into account the different spacings of the observation times. The subject-specific random effects a_i were assumed to follow a normal distribution with mean 0. The estimates were calculated using a Monte Carlo sample size of 1000.

The estimates and their standard errors are given in Table 3. Both analyses give similar results, and the efficiency of the case-cohort design is on average 65% while using half the sample size of the full cohort analysis. The case-cohort design yielded the log odds estimates 0.95 for cats, 0.34 for the second tertile of endotoxin exposure, and 0.41 for the highest tertile. The estimate of the random-effects variances were $\bar{\sigma}^2_\alpha = 0.93$ with a standard error of 0.51 and $\bar{\sigma}^2_g = 3.24$ with a standard error of 0.68. The estimate of γ was 0.85 with a small standard error of 0.05. The log odds estimates based on the full cohort were 0.60 for cats in the house, 0.55 for the second tertile of endotoxin exposure, and 0.44 for the highest tertile. The interaction terms were not statistically significantly different from 0, but indicated that early childhood exposure to endotoxin may have a protective effect at older ages. This is consistent with the scientific literature (see Litonjua et al., 2002). Efficiency of the estimates ranged from 54% for the second tertile of endotoxins to 100% for γ. The average efficiency for the case-cohort design was 66% for half the number of samples.

Table 3
Estimation results of μ, σ_a^2, β, σ_g^2, and γ for children from the Home Allergen Study (with standard errors in parentheses)

Parameter	Cohort analysis	Longitudinal case-cohort design
μ	−4.04 (0.29)	−3.94 (0.37)
Cat in house	0.60 (0.24)	0.95 (0.31)
Endotoxins, first tertile (E1)	0.00 (baseline)	0.00 (baseline)
Endotoxins, second tertile (E2)	0.55 (0.28)	0.34 (0.38)
Endotoxins, third tertile (E3)	0.44 (0.28)	0.41 (0.37)
E1 ∗ (age ≥ 12 months)	0.36 (0.23)	0.25 (0.28)
E2 ∗ (age ≥ 12 months)	−0.16 (0.23)	−0.22 (0.29)
E3 ∗ (age ≥ 12 months)	−0.17 (0.23)	−0.44 (0.30)
σ_a^2	1.71 (0.41)	0.93 (0.51)
σ_g^2	3.06 (0.57)	3.24 (0.68)
ρ	0.74 (0.05)	0.85 (0.05)
Log likelihood	−2083.19	−1209.81

The large estimates of σ_g^2 for both models indicate that intra-individual correlation should not be ignored in the analysis of these data.

6. Discussion

We propose a longitudinal case-cohort design that consists of taking a subsample from the cohort at the beginning of the follow-up period, and a sample of the cases over the course of the study. This design can significantly reduce cost and effort of exposure assessment in epidemiologic cohort studies, while providing encouragingly efficient estimates of measured risk factors. It is most beneficial when a large part of the study cost is in exposure measurement compared to ascertainment of individuals. If the main cost is in ascertainment, a full cohort analysis might be the more sensible approach to analysis. While in the case-cohort design of Prentice (1986) efficiency of the estimates is nearly 100%, the efficiency of our design for log odds estimates was found to be up to 83% with 50% savings in sample size. Our design is slightly less efficient because first, we do not observe all events that occur in the cohort, and second, we are not in the situation of such rare outcomes, where the cases and a reasonable number of controls provide nearly all the information in the sample. The outcomes in the longitudinal settings for which we believe our design to be useful are slightly more common, like asthma attacks, and thus one gains less by oversampling the cases than in a situation of extremely rare outcomes, for example, cancer.

Different correlations between observations on the same individual are accommodated by a two-level random-effects model. One random effect, a, captures between individual variation, and the other random effects, g_t, $t = 1, \ldots, T$, allow intra-individual correlations across time to vary. We introduced a novel application of the bootstrap to test $\sigma_g^2 = 0$ in the presence of unidentifiable nuisance parameters. A simulation study indicated that misspecification of the distribution of underlying individual-level random effects in modeling binary disease outcomes did not lead to serious bias in the model parameters.

We show how our design relates to the bidirectional crossover design by Navidi (1998). The basic difference aside from introducing the control-cohort is that we conditioned the distribution of the cases on the exact ascertainment and allow for different correlations of the observations by using a two-level random-effects model for binary outcomes on the same individual over time. In related work, Dewjani and Moolgavkar (2000) presented a Poisson process

formulation for studying the association between environmental covariates and recurrent events. In the case of a single failure for each subject their likelihood coincides with the likelihood for case-crossover design (Navidi, 1998). Other recent work on environmental covariables has focused on the use of generalized additive models for investigating associations between air pollution measurements and daily counts of events such as hospital admissions (Schwartz, 1994; Moolgavkar, 2000).

A complication of our design that is also shared by the case-cohort or the nested case-control design, is that covariates of the cases only become available after the event of interest has occurred. This makes the design impractical for situations where internal time-varying covariates or time-lagged covariates are relevant. However, in many applications, it is possible to determine at time after failure what an individual's levels of exposure were at the times before failure, for example, using serially stored specimens that can be evaluated when a person has an event. A field of application where such covariate information is available and that was the motivation for our work, is the investigation of air pollutants and respiratory illnesses. Increasing attention is being paid to the potential for urban air toxins and pollutants to exacerbate asthma, through induction of specific or nonspecific airway hyperreactivity, or of reactive airways dysfunction syndrome. Individual hypersensitivity to specific substances may also play a role. A related example is a study, still in the design phase, that addresses the question of prediction of asthma attacks based on knowledge of an individual's susceptibility to pollen. The covariate of interest in this situation is the interaction of sensitivity of a person to a specific pollen and pollen count in an area at a given time. Pollen count data are available from external sources and a skin test is used to determine individual sensitivity. If the longitudinal case-cohort design were used as a study design, one could also estimate the main effect for sensitivity that is constant over time and thus can be estimated neither from the bidirectional crossover design nor from the conditional panel design.

Another example where biospecimens were sampled and stored for later evaluation was described by Park and Kim (2004). For a study of the association of diarrhea and the presence of Enterotoxigenic *Escherichia coli* (ETEC) in children in South Korea, stool samples or rectal swabs were collected daily by the children's parents. However, identification of children with ETEC infection requires costly and time-consuming laboratory testing and the number of diarrhea cases was small compared to the number of repeated measurements. Our case-cohort design would provide a cost-effective way to obtain estimates of the effect of ETEC on diarrhea risk in this setting. Other potential applications of our design include genetic testing of serum samples that are often collected at baseline in cohort studies to assess the impact of gene–environment interactions on repeated outcomes.

Acknowledgements

We thank Diane Gold and Don Milton for access to the data and Mitchell Gail and Matt Wand for helpful comments.

References

Albert, P. S., Follmann, D. A., Wang, S. A., and Suh, E. B. (2002). A latent autoregressive model for longitudinal binary data subject to informative missingness. *Biometrics* **58,** 631–642.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88,** 9–25.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. New York: Springer-Verlag.

Commenges, D. and JacqminGadda, H. (1997). Generalized score test of homogeneity based on correlated random effects models. *Journal of the Royal Statistical Society, Series B* **59,** 157–171.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64,** 247–254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74,** 33–43.

Dewanji, A. and Moolgavkar, S. H. (2001). A Poisson process approach for recurrent event data with environmental covariates. *Environmetrics* **11,** 665–673.

Diggle, P., Liang, K., and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. New York: Oxford University Press.

Gold, D. R., Burge, H. A., Carey, V., Milton, D. K., Platts-Mills, T., and Weiss, S. T. (1999). Predictors of repeated wheeze in the first year of life: The relative roles of cockroach, birth weight, acute lower respiratory illness, and maternal smoking. *American Journal of Respiratory and Critical Care Medicine* **160,** 227–236.

Hartford, A. and Davidian, M. (2000). Consequences of misspecifying assumptions in nonlinear mixed effects models. *Computational Statistics and Data Analysis* **34,** 139–164.

Kupper, L. L., McMichael, A. J., and Spirtas, R. (1975). A hybrid epidemiologic study design useful in estimating relative risk. *Journal of the American Statistical Association* **70,** 524–528.

Langholz, B. and Borgan, O. (1997). Estimation of absolute risk from nested case-control data. *Biometrics* **53,** 767–774.

Liang, K. Y. (1987). A locally most powerful test for homogeneity with many strata. *Biometrika* **74,** 259–264.

Litonjua, A. A., Milton, D. K., Celedon, J. C., Ryan, L., Weiss, S. T., and Gold, D. R. (2002). A longitudinal analysis of wheezing in young children: The independent effects of early life exposure to house dust endotoxin, allergens, and pets. *Journal of Allergy and Clinical Immunology* **110,** 736–742.

Maclure, M. (1991). The case-crossover design: A method for studying transient effects on the risk of acute events. *American Journal of Epidemiology* **133,** 144–153.

Miettinen, O. S. (1982). Design options in epidemiology research: An update. *Scandinavian Journal of Work, Environment, and Health* **8**(suppl. 1), 1295–1311.

Moolgavkar, S. H. (2000). Air pollution and hospital admissions for diseases of the circulatory system in three US metropolitan areas. *Journal of the Air and Waste Management Association* **50,** 1199–1206.

Navidi, W. (1998). Bidirectional case-crossover designs for exposures with time trends. *Biometrics* **54,** 596–605.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* **59,** 25–35.

O'Neill, T. J. and Barry, S. C. (1995). Truncated logistic regression. *Biometrics* **51,** 533–541.

Park, E. and Kim, Y. (2004). Analysis of longitudinal data in case–control studies. *Biometrika* **91,** 321–330.

Pfeiffer, R. M., Gail, M. H., and Pee, D. (2001). Inference for covariates that accounts for ascertainment and random genetic effects in family studies. *Biometrika* **88,** 933–948.

Prentice, R. L. (1986). A case–cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73,** 1–11.

Raudenbush, S. W., Yang, M. L., and Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics* **9,** 141–157.

Schwartz, J. (1994). Air pollution and hospital admissions for the elderly in Birmingham, Alabama. *American Journal of Epidemiology* **139,** 589–598.

Tanner, M. A. (1993). *Tools for Statistical Inference*. New York: Springer-Verlag.

Zhao, L. P. and Lipsitz, S. (1992). Designs and analysis of 2-stage studies. *Statistics in Medicine* **11,** 769–782.

Received May 2004. *Revised October* 2004.
Accepted January 2005.